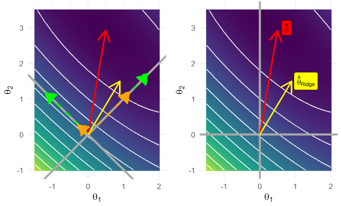
Introduction to Machine Learning

Regularization

Geometry of L2 Regularization



Learning goals

- Have a geometric understanding of *L*2 regularization

GEOMETRIC ANALYSIS OF *L*2 REGULARIZATION

Weight decay can be interpreted **geometrically**.

Let's use a quadratic Taylor approximation of the unregularized objective $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ in the neighborhood of its minimizer $\hat{\theta}$,

$$\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\hat{\theta}) + \nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\hat{\theta}) \cdot (\boldsymbol{\theta} - \hat{\theta}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\theta})^T \boldsymbol{H}(\boldsymbol{\theta} - \hat{\theta}),$$

where $\boldsymbol{H}$ is the Hessian matrix of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ evaluated at $\hat{\theta}$.

- The first-order term is 0 in the expression above because the gradient is 0 at the minimizer.
- $\boldsymbol{H}$ is positive semidefinite, because we are at the minimizer.

GEOMETRIC ANALYSIS OF *L*2 REGULARIZATION / 2

The minimum of $\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta})$ occurs where $\nabla_{\boldsymbol{\theta}}\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta} - \hat{\theta})$ is 0.
Now we *L*2-regularize $\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta})$, such that

$$\tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta}) = \tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

and solve this approximation of $\mathcal{R}_{\text{reg}}$ for the minimizer $\hat{\boldsymbol{\theta}}_{\text{ridge}}$:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}}\tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta}) &= 0, \\ \lambda\boldsymbol{\theta} + \boldsymbol{H}(\boldsymbol{\theta} - \hat{\theta}) &= 0, \\ (\boldsymbol{H} + \lambda\boldsymbol{I})\boldsymbol{\theta} &= \boldsymbol{H}\hat{\theta}, \\ \hat{\boldsymbol{\theta}}_{\text{ridge}} &= (\boldsymbol{H} + \lambda\boldsymbol{I})^{-1}\boldsymbol{H}\hat{\theta}, \end{aligned}$$

This gives us a formula to see how the minimizer of the *L*2-regularized version is a transformation of the minimizer of the unpenalized version.

GEOMETRIC ANALYSIS OF $L2$ REGULARIZATION / 3

- As λ approaches 0, the regularized solution $\hat{\theta}_{\text{ridge}}$ approaches $\hat{\theta}$. What happens as λ grows?
- Because $\boldsymbol{H}$ is a real symmetric matrix, it can be decomposed as $\boldsymbol{H} = \boldsymbol{Q}\Sigma\boldsymbol{Q}^\top$, where Σ is a diagonal matrix of eigenvalues and $\boldsymbol{Q}$ is an orthonormal basis of eigenvectors.
- Rewriting the transformation formula with this:

$$
\begin{aligned}
\hat{\theta}_{\text{ridge}} &= \left(\boldsymbol{Q}\Sigma\boldsymbol{Q}^\top + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{Q}\Sigma\boldsymbol{Q}^\top\hat{\theta} \\
&= \left[\boldsymbol{Q}(\Sigma + \lambda\boldsymbol{I})\boldsymbol{Q}^\top\right]^{-1}\boldsymbol{Q}\Sigma\boldsymbol{Q}^\top\hat{\theta} \\
&= \boldsymbol{Q}(\Sigma + \lambda\boldsymbol{I})^{-1}\Sigma\boldsymbol{Q}^\top\hat{\theta}
\end{aligned}
$$

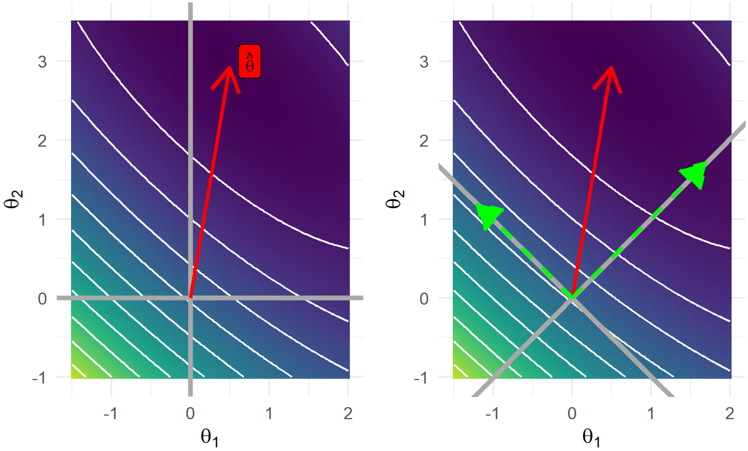
GEOMETRIC ANALYSIS OF *L*2 REGULARIZATION / 4

- Therefore, weight decay rescales $\hat{\theta}$ along the axes defined by the eigenvectors of $\boldsymbol{H}$. The component of $\hat{\theta}$ that is aligned with the *j*-th eigenvector of $\boldsymbol{H}$ is rescaled by a factor of $\frac{\sigma_j}{\sigma_j+\lambda}$, where σ_j is the corresponding eigenvalue.

GEOMETRIC ANALYSIS OF $L2$ REGULARIZATION / 5

Firstly, $\hat{\theta}$ is rotated by $\boldsymbol{Q}^\top$, which we can interpret as a projection of $\hat{\theta}$ on the rotated coordinate system defined by the principal directions of $\boldsymbol{H}$:

GEOMETRIC ANALYSIS OF *L*2 REGULARIZATION / 6

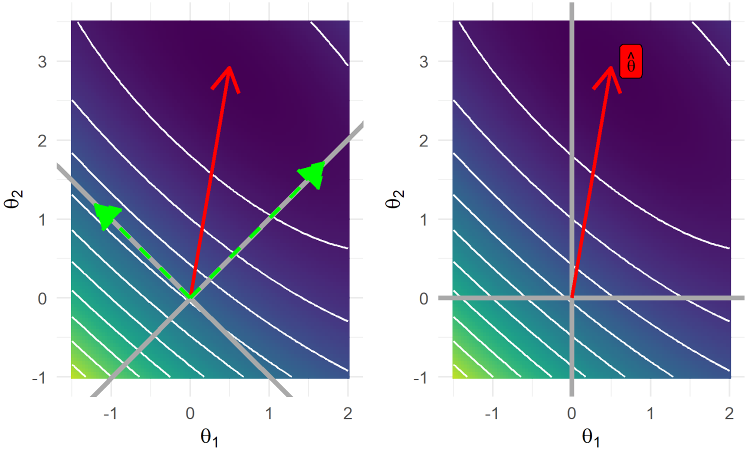


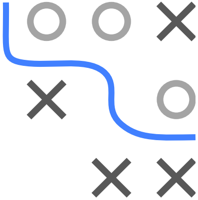
GEOMETRIC ANALYSIS OF $L2$ REGULARIZATION / 7

Since, for $\lambda = 0$, the transformation matrix $(\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} = \boldsymbol{I}$, we simply arrive at $\hat{\theta}$ again after projecting back.

GEOMETRIC ANALYSIS OF *L*2 REGULARIZATION

/ 8



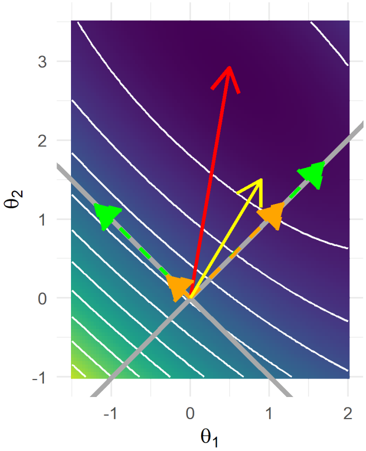


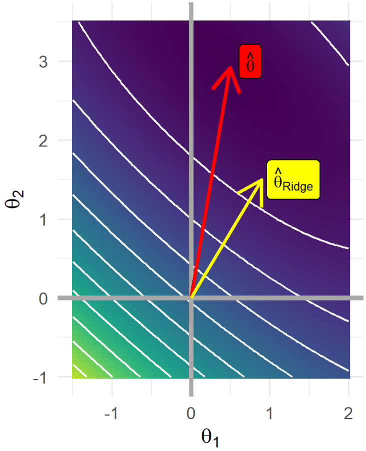
GEOMETRIC ANALYSIS OF $L2$ REGULARIZATION / 9

If $\lambda > 0$, the component projected on the j-th axis gets rescaled by $\frac{\sigma_j}{\sigma_j+\lambda}$ before $\hat{\theta}_{\text{ridge}}$ is rotated back.

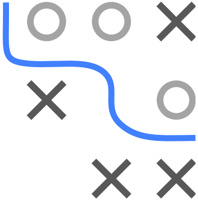
GEOMETRIC ANALYSIS OF *L*2 REGULARIZATION

/ 10





GEOMETRIC ANALYSIS OF *L*2 REGULARIZATION



- Along directions where the eigenvalues of $\boldsymbol{H}$ are relatively large, for example, where $\sigma_j >> \lambda$, the effect of regularization is quite small.
- On the other hand, components with $\sigma_j << \lambda$ will be shrunk to have nearly zero magnitude.
- In other words, only directions along which the parameters contribute significantly to reducing the objective function are preserved relatively intact.
- In the other directions, a small eigenvalue of the Hessian means that moving in this direction will not significantly increase the gradient. For such unimportant directions, the corresponding components of $\boldsymbol{\theta}$ are decayed away.

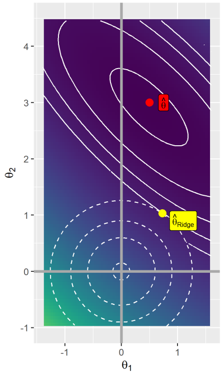


Figure: The solid ellipses represent the contours of the unregularized objective and the dashed circles represent the contours of the *L*2 penalty. At $\hat{\boldsymbol{\theta}}_{\text{ridge}}$, the competing objectives reach an equilibrium.